

工作周报

05.21.2012 - 05.27.2012

马昱欣

本周小结

本周完成了毕业论文写作。接下来的重点主要放在教材编写和信任关系可视化项目上。

论文阅读

本周主要阅读了Jeffrey Heer小组在AVI 2012上的两篇文章《GraphPrism: Compact Visualization of Network Structure》和《Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment》，两篇文章均与近期该小组发布的Datavore工具相关。同时本周还浏览了龙星计划课程上Jiawei Han的Social Network Analysis报告，主要是近年来社交网络研究方面的内容概述。

科研进展

受到近期淘宝各种“差评门”的启发，想到当前项目中对信任数据的使用不够深入，抛弃了评分时间等属性，会对整个数据中的pattern抽取有较大影响。现在的一个拓展思路是在信任评分数据的基础上融合其他类型的数据，比如评分时的用户留言、评分时间等信息，以发现类似淘宝中“职业差评师”一类的用户。并且在需求方面，不知道淘宝是否有这样类似的需求。

工程进展

本周对淘宝开发的OceanBase做了一下大致了解。当前OceanBase的文档比较少，主要是淘宝开发团队提供的一些软件文档，比较缺乏用户体验、使用经验等类型的文章。从安装配置上讲，OceanBase较其他开源产品来说要繁琐得多，当前在Mac平台上还未成功运行。

在网上查找资料的过程中，看到一篇讲HBase相关内容的文章，其中的一句话比较赞同：大数据时代，数据存储没有银弹可用。因此在当前可视分析平台上搭建数据存储管理系统的话，需要使用多种数据库来应对不同的数

据类型，之后增加中间件来统一异构数据源之间的关系。不过是否需要将过多精力投入在这一部分上确实需要仔细斟酌。

在分析过程上有一点思考是，是否可以像Voreen、Orange等软件一样，将分析功能模块化，使得功能与功能之间能够组成pipeline。这是否涉及一些可视分析原语一类的内容，值得下一步深入思考。

教材进展

周四浏览Social Network Analysis的时候发现Jiawei Han的数据挖掘教材出了2011新版，其中对数据挖掘部分的内容做了很大的修订，打算在编写教材时参考这本新版教材。数据清洗、去噪部分主要读了《Research directions in data wrangling: Visualizations and transformations for usable and credible data》这篇文章。

下周计划

- (1) 继续整理教材相关资料，开始动手修改原始稿
- (2) 基于本周的启发点对信任关系可视化项目进行拓展思考
- (3) 准备周三组会报告、周六毕业论文答辩